Microevolutionary traits and comparative population genomics of the emerging pathogenic fungus Cryptococcus gattii

Farrer, Rhys A.; Voelz, Kerstin; Henk, Daniel A.; Johnston, Simon; Fisher, Matthew C.; May, Robin C.; Cuomo, Christina

DOI:
10.1098/rstb.2016.0021

PHILOSOPHICAL TRANSACTIONS B

Microevolutionary traits and comparative population genomics of the emerging pathogenic fungus Cryptococcus gattii

Journal:	*Philosophical Transactions B*
Manuscript ID	RSTB-2016-0021.R1
Article Type:	Research
Date Submitted by the Author:	n/a
Complete List of Authors:	Farrer, Rhys; Broad Institute, Genome Sequencing and Analysis Program; Imperial College London, Infectious Disease Epidemiology Voelz, Kerstin ; University of Birmingham, Institute of Microbiology and Infection & School of Biosciences; University Hospital Birmingham NHS Foundation Trust, NIHR Surgical Reconstruction and Microbiology Research Centre Henk, Daniel; Imperial College London, Infectious Disease Epidemiology Johnston, Simon; University of Birmingham, Institute of Microbiology and Infection & School of Biosciences Fisher, Matthew; Imperial College London, Infectious Disease Epidemiology May, Robin; University of Birmingham, Institute of Microbiology and Infection & School of Biosciences; University Hospital Birmingham NHS Foundation Trust, NIHR Surgical Reconstruction and Microbiology Research Centre Cuomo, Christina; Broad Institute, Genome Sequencing and Analysis Program
Issue Code: Click here to find the code for your issue.:	FUNGAL
Subject:	Evolution < BIOLOGY, Genomics < BIOLOGY, Microbiology < BIOLOGY
Keywords:	Cryptococcus gattii, microevolution, mitochondrial tubularisation, intracellular proliferation

Microevolutionary traits and comparative population genomics of the emerging pathogenic fungus *Cryptococcus gattii*

Rhys A. Farrer[1,2,*,†], Kerstin Voelz[3,4,*], Daniel A. Henk[2], Simon A. Johnston[3,5], Matthew C. Fisher[2], Robin C. May[3,4], Christina A. Cuomo[1]

[1]Genome Sequencing and Analysis Program, The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA.
[2]Dpt. of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, UK
[3]Institute of Microbiology and Infection & School of Biosciences, University of Birmingham, Birmingham, UK
[4]NIHR Surgical Reconstruction and Microbiology Research Centre, University Hospitals Birmingham NHS Foundation Trust, Queen Elizabeth Hospital Birmingham, Birmingham, UK
[5]Current address: Department of Infection, Immunity and Cardiovascular Disease, Medical School, University of Sheffield, Sheffield, UK.
[*]These authors contributed equally to this work.
[†]Corresponding author

Abstract

Emerging fungal pathogens cause an expanding burden of disease across the animal kingdom, including a rise in morbidity and mortality in humans. Yet, we currently only have a limited repertoire of available therapeutic interventions. A greater understanding of the mechanisms of fungal virulence, and the emergence of hypervirulence within species are therefore needed for new treatments and mitigation efforts. For example, over the last decade, an unusual lineage of *Cryptococcus gattii*, which was first detected on Vancouver Island, has spread to the Canadian mainland and the Pacific Northwest infecting otherwise healthy individuals. The molecular changes that led to the development of this hypervirulent cryptococcal lineage remain unclear. To explore this, we traced the history of similar microevolutionary events that can

lead to changes in host-range and pathogenicity. Here, we detail fine-resolution mapping of genetic differences between two highly-related *Cryptococcus gattii* VGIIc isolates that differ in their virulence traits (phagocytosis, vomocytosis, macrophage death, mitochondrial tubularisation, and intracellular proliferation). We identified a small number of single site variants within coding regions that potentially contribute to variations in virulence. We then extended our methods across multiple lineages of *C. gattii* to study how selection is acting on key virulence genes within different lineages.

Keywords

Cryptococcus gattii, microevolution, mitochondrial tubularisation, intracellular proliferation

Introduction

Emerging fungal pathogens and fungal-like organisms are increasingly threatening natural populations of animals and plants [1]. For example, the recently discovered chytrid fungus *Batrachochytrium salamandrivorans* was implicated in the near extirpation of fire salamanders in 2013 in the Netherlands [2]. Race Ug99 of the basidiomycetous fungus *Puccinia graminis* f. sp. *tritici* first detected in 1998 is now recognized as a major threat to wheat production and food security worldwide [3], and the basidiomycetous fungus *Cryptococcus gattii* (*C. gattii*) has expanded its range into non-endemic environments with a consequential increase of fatal meningitis in humans [4,5]. The global threat of these and other related diseases is underpinned by fungi harbouring complex and dynamic genomes [6]. This leads to an ability to rapidly evolve in order to overcome host-defences [7], presenting a pressing challenge to understand the mechanisms that drive the evolution of the phenotypic determinants that underlie pathogenicity.

C. gattii causes pneumonia and meningoencephalitis in humans following inhalation of infectious yeast or airborne hyphae [8]. While its sister species *C. neoformans* is most prevalent in HIV-infected individuals and patients with other immunodeficiencies, *C. gattii* predominantly (although not

exclusively, e.g. [9]) causes disease in healthy people [10]. *C. gattii* accounts for less than 1% of all cryptococcosis cases and until the late 1990s was found mostly in tropical and subtropical parts of the world. However, in 1999, an outbreak of *C. gattii* was reported on Vancouver Island in domestic pets, wild animals, and people [4,11]. This outbreak spread to mainland Canada and then into the Northwestern states of the United States and remains a major public health concern.

C. gattii is divided into four distinct lineages (VGI, VGII, VGIII and VGIV), with such considerable genetic variation that they were recently described as separate species (*C. gattii*, *C. deuterogattii*, *C. bacillisporus* and *C. tetragattii* respectively [12]). VGI and VGII isolates are responsible for the majority of infections in immunocompetent individuals in the Pacific Northwest, the North of Australia, and in Central Papua New Guinea [13]. Although the original outbreak on Vancouver Island was caused by at least two clonal subgroups of VGII named VGIIa (the major genotype) and VGIIb (the minor genotype) [11], several associated outbreaks have subsequently been reported, e.g. VGIIc in Oregon, United States [14]. Recent studies investigating the genetic diversity of outbreak isolates by whole genome sequence typing have identified an abundance of genetic diversity within the VGII molecular type and evidence for both sexual recombination and clonal expansions [15–17].

The ability of cryptococcal cells to parasitise phagocytes, in particular macrophages, is a major pathogenesis mechanism of cryptococcosis [18,19]. *C. gattii* is able to protect itself from host induced oxidative stresses, such as reactive oxygen species (ROS) *via* an enlarged polysaccharide capsule, which provides a physical barrier that interferes with normal macrophage phagocytosis and clearance by the immune system [20]. Although all four lineages are capable of causing disease, a number of differences have been identified between sub-lineages, such as increased intracellular proliferation rates (IPR) in VGIIc isolates [5], or an enhanced ability to parasitise host phagocytic cells by VGIIa outbreak isolates. These processes are initiated upon engulfment by macrophages, followed by a stress response that triggers

cryptococcal mitochondrial tubularisation and rapid proliferation of the outbreak strains [19]. Another study identified increased expression levels for laccase in the VGIIa isolate R265 compared with non-outbreak strains; laccase controls melanin production, and provides protection from oxidative damage imposed by the host immune response. In addition, cryptococcal strains are able to escape phagocytes by a non-lytic mechanism (expulsion or 'vomocytosis' [21,22]) or to undergo 'lateral transfer' between phagocytes. These processes may provide greater resistance to stresses in the phagosome and may also have a role in the dissemination of the pathogen from the lungs to the central nervous system.

Genomic comparisons between lineages have identified a range of genetic differences that may contribute to differences in fitness, ranging from chromosome copy number variation to genomic rearrangements [23,24]. Furthermore, as many as 700 genes are unique to one or more of the four lineages, including heat shock proteins and iron-transporters, which could contribute to differences in disease progression and outcome [24]. Positive selection has also been identified among orthologous multi-drug transporters in different lineages and clonal-groups [24]. However, new emerging or hypervirulent genotypes arise at the population level, and may not be detected from comparisons between more anciently diverged isolates. Here, we combine phenotypic typing from 20 *C. gattii* strains from each of the four lineages, with a case-study genomic comparison for two highly genetically similar isolates belonging to VGIIc (EJB18 and EJB52) that have marked differences in intracellular proliferation rates (IPR) and mitochondrial tubularisation rates. Our approach identifies 33 candidate nuclear genes that may contribute to these hypervirulent traits. Finally, we extend our approach to study the wider population structure and variation among a panel of 64 *C. gattii* isolates and demonstrate how the methods we detail here are applicable to investigating the genetic determinants that underpin virulence across a wide range of emerging fungal pathogens.

Results

Whole genome sequencing and phenotypic typing suggests a loss/gain pattern of hypervirulence traits among VGIIc

Detecting micro-evolutionary changes requires precision variant-calling to distinguish subtle differences with sequencing or alignment error. Using a highly stringent SNP-calling protocol (**see methods**), we were able to reduce false positive SNPs to 4.5 per 1 million bases (*n*=77; 0.13% of all SNPs called), while maintaining 99.3% true positive SNPs (the remaining 0.7% were false negatives). Applying these parameters to 66 isolates of *C. gattii* (**Tables 1 and S1**), we identified SNPs in all isolates and used these to construct a phylogenetic tree, illustrating the sub-lineages within each of the four major lineages of *C. gattii* (**Figure 1**).

Macrophage parasitism and the ability to proliferate within these phagocytic effector cells are well-established virulence traits of cryptococcal infections. To correlate genetic and phenotypic distance, we measured a range of macrophage interaction parameters (i.e. phagocytosis, intracellular proliferation, vomocytosis, cryptococcal mitochondrial tubularisation, and macrophage death) in replicate (3X-8X) over a timecourse of macrophage interaction (0h, 18h, 24h, 48h). First, we measured the maximal intracellular proliferate potential (T_{max}; commonly referred to as intracellular proliferation rate; IPR) for 20 isolates of *C. gattii* including four of the six named subclades (the outbreak clades VGIIa and VGIIc, the recently described VGIIx [24], and VGIIb) (**Figure 1 and Table S2**). These strains were selected, according to previous literature and strain detail knowledge, to represent a balanced collection of strains i) from the North Pacific *C. gattii* outbreak, ii) from environmental origin and iii) representing the different molecular groups. IPR values ranged from 0.74 to 2.30, and K-means clustering revealed 2 groups. One group contained isolates with lower values found across the all the lineages and therefore not correlated with phylogeny (all less than 1.5 IPR). These included one VGI isolate, 5 of the 7 non-clonal subclades of VGII, both VGIIIb isolates, and one VGIV isolate. In contrast, all six of the VGIIa and the VGIIx isolate were in the cluster with higher values (greater than 1.7 IPR).

In addition to IPR, we measured host-pathogen interactions *via* the induction of mitochondrial tubularisation (specifically, an average percent of yeast with tubular mitochondria), which is associated with response to, and protection from, reactive oxygen species (ROS) upon engulfment [25]. Again, values were highly variable among the four lineages (**Table S2**) – with the highest proportion of tubularising mitochondria among the subclades of VGII. Intracellular proliferative capacity correlated significantly with mitochondrial tubularisation and yeast uptake by macrophages ($p < 0.0001$ and $p = 0.004$, respectively, **Figure 1**) [25]. Of note, two VGIIc isolates show large differences in mitochondrial tubularisation (EJB52 = 18%, and EJB18 = 44%) and IPR (EJB52 = 1.36 ± 0.33 IPR, and EJB18 = 1.71 ± 0.35 IPR), suggesting these isolates are suitable for further investigation. Substantial variation was also observed in the fraction of yeast phagocytised, with uptake percent varying from 0.5% to 31.5% (**Figure 1, Table S2**). The greatest percent uptakes were from most of VGIIa, VGIIx and importantly VGIIc EJB52 (28.4%), while VGIIc EJB18 was phagocytised less (15.7%).

We also assayed each isolate for the rate of expulsion after phagocytosis by macrophages ('vomocytosis') and macrophage cell death. In contrast to IPR, mitochondrial tubularisation and percentage phagocytosis, there was far lower intra-lineage variation in non-lytic escape ('vomocytosis') and in host cell death (**Figure 1, Table S2**). Furthermore, neither of these two phenotypic markers significantly correlated ($p = 0.837$ and $p = 0.235$, respectively) with intracellular proliferation rate. This may be an indication that *C. gattii* hypervirulence is driven by the correlated phenotypes and not by 'vomocytosis' or macrophage death.

Comparing closely related isolates for shifts in phenotypes (the pattern of IPR, mitochondrial tubularisation, and phagocytosis) shows the greatest discrepancy between the two VGIIc isolates EJB52 and EJB18. A two-tailed *t*-test of the IPR replicates also suggested a difference in values (*p*-value = 0.0458). The limited genetic variation and large phenotypic difference between these two isolates suggested they were good candidates to identify the genetic differences that may be responsible.

To determine if phenotypic variation in VGIIc stemmed from gene loss, we measured read coverage and depth across each gene. Read depth revealed a total of 686 presence/absence (P/A) polymorphisms in at least one isolate, of which 16 were absent in both VGIIc isolates (**Table S3**). No P/A polymorphisms were found in VGIIa isolates to which the reference R265 belongs. Although the 16 P/A polymorphisms in VGIIc could not explain the phenotypic differences between the two VGIIc isolates, they may be relevant to phenotypic differences between subclades (such as between VGIIa and VGIIc). Three of the VGIIa genes absent in VGIIc have both zinc-binding dehydrogenase and alcohol dehydrogenase GroES-like PFAM domains. These are likely the most abundant zinc-binding proteins in the cell, and may play a role during zinc-deprivation conditions such as within a phagosome [26]. Two additional proteins had Major Facilitator Superfamily and sugar transport PFAM domains, which may play out as differences in the ability to transport toxins or xenobiotics out of the cell, or transport sugars into the cell, respectively.

We classified every base of the 17.3 Mb genomes of EJB18 and EJB52, placing 98.65% into a non-ambiguous sub-category (the remaining 1.35% of the genomes were ambiguous i.e. poorly supported base call in one or both of the two isolates). Nucleotide sub-categories were annotated in a codon-by-codon manner, as either fixed or transitional between the two isolates. Only 153 differences (8.9 per Mb; **Figure 2A, Table S4**) were identified between these isolates (compared with 60 thousand; 3.5 per kb from their initial alignments to the VGIIa R265 reference genome; **Table S1**). Importantly, no differences were detected within the mitochondrial genome, suggesting the phenotypic differences in the regulation of mitochondrial tubularisation are encoded in the nuclear genome. Furthermore, no aneuploidy was detected based on depth of coverage plots.

Of the 153 sites differing between EJB18 and EJB52, a subset of 33 variants overlapped sites among the 6,456 predicted genes of R265, including 15 SNPs and 18 indels (**Figure 2C**). The remaining differences fell within

introns (*n*=32) and intergenic (*n*=88) regions, differences that could also have an impact if they, for example, fell within a promoter or repressor region. To examine this, we identified all intergenic differences that were upstream of a gene (**Figure 2B**), 13 of which were within 100nt. The closest upstream intergenic difference was an insertion unique to EJB52, 12 bases upstream of the STE/STE11/SSK protein kinase (CNBG_4621), involved in *Cryptococcus* mating and virulence [27]. Separately, a solute carrier family 25 (mitochondrial citrate transporter) gene had 2 upstream intergenic differences: an insertion 46nt upstream unique to EJB52 and a deletion 44nt upstream unique to EJB18. Improper uptake or conversion of citrate is known to attenuate a range of pathogens, including *Cryptococcus* [28–30].

Of the 31 genes that had differences between EJB52 and EJB18 (**Figure 2C**) (2 genes had 2 differences), 2 were synonymous changes, and 8 were hypothetical proteins without any assigned functional information (GO-terms, PFAM, TIGRFAM). The remaining 23 genes had non-synonymous differences and functional annotation. As mitochondrial tubularisation is one of the main phenotypic differences between the 2 VGIIc isolates, we predicted proteins that are localised to the mitochondria. Of the 6,456 proteins in *C. gattii* VGII, only 548 were predicted to localise to the mitochondria; 2 of these genes had genetic differences between the 2 VGIIc isolates. The first gene (CNBG_5651) has an insertion (Non-frameshift / modulo 3) in EJB18 (high value tubularisation), and its specific function is unclear. Furthermore, this allele is not correlated with the mitochondrial tubularization phenotype across the isolates as the insertion was found in 40 of 66 isolates including both high and low mitochondrial tubularisation percentages. The second gene (CNBG_5312) has a non-synonymous change (T->C, I->V) in glucoamylase in EJB18; this change is unique to EJB18. Glucoamylase expression is responsible for carbohydrate metabolism in other intracellular pathogens, such as *Listeria pneumophila* within amoebae [31,32], for which *C. gattii* is also likely evolving for protection against [33].

Although not predicted to localise to the mitochondria, a gene with the mitochondrial carrier protein domain (CNBG_4812) has a unique synonymous

SNP in EJB52 (low tubularisation) compared with EJB18; this change also does not fall within a splice donor/acceptor site. Curiously, this SNP was also present in 6 of the 7 VGI isolates sequenced, including WM276 that also shows low tubularisation. The other strains with this change were not measured for this phenotype, including the 6 other VGI isolates and one other VGIIc isolate (B7466). Finally, it is noteworthy that two separate histone deacetylases vary between the isolates, including a synonymous SNP unique to EJB52 in CNBG_3873 annotated as '6/10/Arb2' and a deletion in EJB52 -> insertion in EJB18 for CNBG_1847 annotated as 'RPD3'. The deletion was found in only 1 other isolate: the closely related VGIIc B7466, for which we have no phenotypic data. However, it is possible that unique variants give rise to shared phenotypes by disrupting the same gene or gene network. For example, histone deacetylases are involved in the morphology and virulence of *C. albicans* [34].

Selection in the clonal groups of *C. gattii* acts upon capsule genes, heat-shock proteins and the STE/STE11/SSK protein kinase

To examine the impact of selection on the different *C. gattii* clades, we measured d_N/d_S (ω) values for fixed differences across 6 subclades of *C. gattii* (2a, 2b, 2c, 2x, 3a, 3b). A total of 859 genes have d_N/d_S with values greater than 1, which can be indicative of relaxed or positive selection (**Fig. 3**). No PFAM or GO-term from these genes were enriched (according to two-tailed Fisher's exact test with *q*-value FDR against the remaining gene-sets). However, two genes of known interest were identified. The first gene (CNBG_1370) is a 1:1 ortholog of *C. neoformans* H99 Utr2 gene (with homology to chitin transglycolase), which is potentially involved in capsule biosynthesis [35]. In *C. gattii* this gene is under selection in the recently named VGIIx subclade [24] with d_N=0.0063, d_S=0.0059, ω=1.0792. The second gene (CNBG_0047) is in an orthogroup with *C. neoformans* H99 Cap64-like proteins Cas31 and Cas3, and under selection in VGIIb (d_N=0.0028, d_S=0.0023, ω=1.2242). Both Cas31 and Cas3 mutants have decreased capsule sizes in *C. neoformans*, as they are involved in determining the position and the linkage of the xylose and/or O-acetyl

residues on the mannose backbone of the capsule [35,36]

The reference isolate VGIIa R265 is only separated from other VGIIa isolates by between 39 and 184 SNPs (**Table 2**), of which only 12 were fixed (i.e. likely errors in the R265 assembly itself). To measure selection in VGIIa we therefore used the branch site model (BSM) of selection implemented by PAML [37]. We found 113 genes with significant differences ($\chi^2_2(2\Delta\text{Lnl}) < 0.01$) after Benjamini Hochberg (BH) multiple correction. Previously when we calculated these values for VGIIa (Subset 5 in [24]) for non-fixed differences across multiple isolates with BH correction and *q*-value < 0.01, without adjusting the NSsites parameter, we identified an almost entirely distinct list of 87 genes, apart from 2 genes that were identified in both (CNBG_5460 (Fungal Zn(2)-Cys(6) binuclear cluster domain) and CNBG_4219 (Domain of unknown function 1708)). Again, no PFAM or GO-term from these genes were enriched (according to two-tailed Fisher's exact test with *q*-value FDR). However, we did identify a transcription factor from the Zn(2)-Cys(6) family named CTA4 that is a nitric oxide-responsive element (NORE), and required for the nitrosative stress response in *C. albicans* [38].

Other important genes found to be under selection in the VGIIa branch using the BSM included a Heat Shock Protein 71 (CNBG_5963) and an ABC transporter (MDR/TAP) member 1 (CNBG_9005), both of which are known to be involved in virulence by a range of pathogens (e.g. [39–42]). Selection was found in the VGIIa branch for the STE/STE11/SSK protein kinase (CNBG_4621), which was also identified and discussed earlier for having the closest (12nt) upstream intergenic difference between VGIIc isolates EJB52 and EJB18. Finally, an ortholog to the *C. neoformans* H99 capsule gene Pmt4 (*C. gattii* gene CNBG_0576) was found to be under selection ($\chi^2_2(2\Delta\text{Lnl}) = 6.64\text{E-}06$) – specifically on a histidine at position 264. Futhermore, *C. neoformans* Pmt4 mutants have decreased capsule sizes [35]. Therefore, at least 3 of the 4 clonal subgroups of *C. gattii* VGII (a, b and x) have evidence for selection within microevolutionary time scales in one of their capsule biosynthesis genes.

Discussion

The aetiological agents of infectious disease impose a huge burden on human society. By understanding their biology, reproduction and mechanisms of infection, we are able to assess and discover new strategies for mitigating their impact. In recent years, fungi have gained widespread attention for their ability to rapidly emerge and threaten both animal and plant species across a global scale [1]. However, many features of their genomes that enable them to successfully adapt to infect diverse hosts and inhabit a wide range of ecological niches remain cryptic [6], especially for newly evolved emerging lineages. The underlying mechanism driving outbreaks caused by *Cryptococcus gattii* has been puzzling researchers for over a decade. Compounding this, is that its virulence is likely a consequence of adaptations that have evolved for protection against environmental predators such as amoebae [33] (unless it's surviving and escaping out of dead animals).

Here we have combined whole genome sequencing with phenotypic analysis to identify recent genetic changes that might underpin cryptococcal hypervirulence. Our phenotypic analysis demonstrated that intracellular proliferation and mitochondrial tubularisation, but not phagocytosis or expulsion by the host, correlate strongly with hypervirulence. This finding was unexpected, and may be informative for their predictive potential in future studies working with these phenotypes. One possible explanation is that there is a set of general virulence factors that allow both *C. neoformans* and *C. gattii* to establish within the human host. However, in the *C. gattii* hypervirulent strains, only the subset responsible for intracellular proliferation and mitochondrial tubularisation provide enhanced parasitism of innate immune effectors. Thus, our study supports previous findings that indicate the importance of intracellular proliferation in cryptococcal virulence and suggests that intracellular proliferation (IPR) and mitochondrial tubularisation may be useful as ‘proxy’ markers of virulence. We also found that these features can be surprisingly variable, even within very closely related isolates of a given

subclade, such as those between EJB52 (low tubularisation and IPR) and EJB18 (high tubularisation and IPR) of VGIIc.

No differences were found between the mitochondrial genomes of the VGIIc isolates EJB52 and EJB18, suggesting that differences in their mitochondrial tubularisation stems from differences in their nuclear genome – which already is thought to control mitochondrial function, fusion and fission [43]. In contrast, comparing the nuclear genomes of the VGIIc isolates EJB52 and EJB18 revealed a set of 31 genes that might be involved in the differences we found in cryptococcal hypervirulence. It is unlikely that all of these loci will be directly involved in virulence regulation. However, one or more of these genes are clear candidates, such as the two genes that localise to the mitochondria, including the glucoamylase or the two histone deacetylases. Although the variants were not correlative across the species at large, there may be numerous unique genetic routes to similar phentoypes. Some of the phenotypic differences may also be explained by non-direct, and/or epistatic, pleiotrophic or epigenetic means. Potentially non-direct acting variants were identified within the intergenic regions, sometimes falling close upstream of transcription start sites, such as the insertion unique to EJB52 immediately upstream of the STE11/SSK protein kinase. That this gene appears to be under selection in the sister sub-clade VGIIa suggests that this is at least a region of dynamic variation, if not potentially involved in some phenotypic differences between these isolates.

The approach we have taken in this study is to combine phenotypic screening with whole genome comparisons in an attempt to identify the genomic determinants underpinning fungal virulence. To reduce the impact of sequencing errors, assembly errors in the reference sequence, alignments errors, and variant call errors, it is essential to assess false discovery rates and respond to those sources of error in an iterative approach [44]. In this study, we have identified 153 high confidence genetic differences that could explain differences in virulence traits between isolates. Detection of microevolutionary events can ascribe new mechanisms behind increased virulence, such as the increased expression of FRE3-encoded iron reductase

in *C. neoformans* passaged in mice [45], or a single *de novo* heterozygous position within a gene called SSN3 capable of restoring filamentation in a nonfilamentous *Candida albicans* mutant [46].

The mechanisms behind changes in mitochondrial tubularisation and intracellular proliferation (IPR), which appear to be linked, may be the result of one or more genetic differences between high and low value isolates (such as those identified here). It is from a sufficiently small number of genetic differences, that hypothesis can be generated, and experimentally validated *via* gene disruptions or allele-swaps [47]. Alternatively, large panels of isolates can be screened *via* a genome-wide association study (GWAS) approach [48]. Where protein structures have been resolved, and are available, sites of positive selection are often in regions at the host-pathogen interface [49], further guiding a functional prediction. Finally, upstream variants that impact expression levels could be characterized using RNA-Seq. Ultimately, progress in pathogenomics heavily relies on open access and usability of well maintained databases of sequence data, functional information and annotation, and pathogen specific online resources for community driven efforts.

We complimented the comparison of 2 closely related isolates with selection analysis across fixed variants, for which d_N/d_S ratios were originally developed [50]. By focusing only on fixed differences, we found evidence for selection in capsule biosynthesis genes in each of the other 3 subclades of VGII, each of which had been phenotyped *via* mutants as having an effect on capsule size or likely to be involved in its assembly. Selection across these genes suggest that each subclade of VGII is generating new alleles and variation within the capsule genes, some of which may result in new peaks in an adaptive landscape, and become a distinguishing genetic feature of their clonal expansion.

Material and Methods

Yeast and mammalian cells, growth conditions and phenotypic analysis

Twenty of the sixty-six *Cryptococcus gattii* strains (**Table 1; SRP017762**) typed in this study were cultured in liquid or agar YPD media (1% peptone, 1 %yeast extract, 2% D-(+)-glucose) for 24 h at 25°C rotating at 20 rpm [13,14,19,25,51] prior to experiments. Mammalian J774 macrophage-like cells were grown as described previously [13,14,19,25,51].

Macrophages were infected with yeast cells and intracellular proliferation monitored as previously described [13,14,19,25,51]. Cryptococcal mitochondrial morphology was determined as described previously [5]. We would like to note that some of the IPR and mitochondrial tubularisation data was previously presented (see [25]), although this paper included additional biological repeats. For analysis of vomocytosis and macrophage cell death of sequenced strains time lapse images were captured on a TE2000 (Nikon) enclosed in a temperature controlled and humidified environmental chamber (Okolabs) with 5% CO_2 at 37°C with Digital Sight DS-Qi1MC camera (Nikon), 20x objective (Ph1 PLAN APO), using NIS elements AR software (Nikon). Images were captured every 2 minutes for 24 hours. Vomocytosis (non-lytic expulsion of intracellular cryptococci) and infected macrophage cell death (disintegration of macrophage containing one or more cryptococci) were scored blind from 4 separate experiments for each of the 20 strains. Clusters of phenotypes were inferred via k-means clustering in R (kmeans) with 1000 iterations.

Variant calling and sequence analysis

Alignment and SNP calling parameters were initially optimized. The recently updated [24] nuclear and mitochondrial genome sequences and feature files for *C. gattii* isolate VGIIa R265α were used (GenBank project accession number AAFP01000000). Additional isolates sequenced and described in previous studies [16,24,52] were obtained from the Short Read Archive (SRA) and converted from SRA format to FASTQ using the SRAtoolkit v2.3.3-4. Illumina reads were aligned to the genome sequence using Burrows-Wheeler Aligner (BWA) v0.7.4 mem [53] with default parameters, obtain high depth alignments (average 116X), and converted to pileup format using Samtools v.0.1.18 [54]. To act as a control for sequencing,

alignment and SNP calling, we included the reference strain R265 in our panel of isolates.

The Genome Analysis Toolkit (GATK) [55] v2.7-4-g6f46d11 was used to call both variant and reference bases from the alignments. First, the Picard tools [56] AddOrReplaceReadGroups, MarkDuplicates, CreateSequenceDictionary and ReorderSam were used to preprocess the alignments. We used GATK RealignerTargetCreator and IndelRealigner for resolving misaligned reads close to indels on parental-progeny pairs of isolates to avoid discrepancies between isolates. Next, GATK UnifiedGenotyper (with haploid genotyper ploidy setting) was run with both SNP and INDEL genotype likelihood models (GLM). We additionally ran BaseRecalibrator and PrintReads for base quality score recalibration on those initial sites for GLM SNP and then re-called variants with UnifiedGenotyper (emitting all sites). We next merged and sorted all of the calls, and ran VariantFiltration with the parameters “QD < 2.0, FS > 60.0, MQ < 40.0”. Next, we removed any base that had less than a minimum genotype quality of 50, or a minimum depth of 10. Finally, we removed any positions that were called by both GLMs (i.e. incompatible indels and SNPs), any marked as “LowQual” by GATK, nested indels, or sites that did not include a PASS flag.

To assess the ability of GATK v2.7-4 UnifiedGenotyper to identify variants, we realigned reads from the reference isolate R265 back to the R265 genome after introducing 60,000 SNPs (corresponding to within VGII variation) and calculated the false discovery rate (FDR) [44]. Our alignment and SNP-calling approach were optimised for maximum specificity, which was necessary for characterising microevolutionary differences. Specifically, we identified 59,578 (99.30%) true positive SNP’s, while only finding 77 (0.13%) false positive SNPs. For gene presence/absence polymorphisms, we counted all genes that had <3X depth of coverage.

For our phylogenetic analysis we extracted all positions that were called single base homozygous (reference or SNP) and polymorphic in ≥1 isolate in the 66 isolates (**Fig. 1**) encompassing 1,192,514 nuclear sites and

767 mitochondrial sites. We inferred the phylogeny of the isolates using RAxML v7.7.8 with the GTRCAT model and 1,000 bootstrap replicates. For the PAML [37] selection analysis, we used the same tree building parameters on a subset of variants that were fixed in each of the isolates in one of six subclades, encompassing 647,792 sites.

Genes that localised to the mitochondria were identified using TargetP [57]. For our selection analysis, we calculated d_N/d_S with yn00 of PAML [37] implementing the Yang and Nielsen method [58] on every gene in each of the six subclades (2a, 2b, 2c, 2x, 3a, 3b) using only fixed differences. For VGIIa, we used Codeml of PAML [37], implementing the Branch Site Model (BSM) A (model=2, NSsites=2, fix_omega=0) compared with the null model (model=2, NSsites=3, fix_omega=1, omega=1) on every gene. Next, we calculated a chi-squared test with 2 degrees of freedom for 2 * the log likelihood difference between the two compared models ($\chi^2_2(2\Delta\text{Lnl})$) with Benjamini Hochberg (BH) multiple correction, and significance set at $q < 0.01$. 714 genes had values ranging from 1 to 2.25^{-37}, while the remaining genes did not have values (e.g. due to insufficient genetic distance).

Acknowledgements

We would like to thank Arturo Casadevall for providing the 18B7 antibody used in this study, and Hannah Larner for the genomic library preparation. This work was financially supported by a Lister Fellowship to Robin C. May, the Medical Research Council (G0601171), the Wellcome Trust (WT088148MF) and the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) ERC Grant Agreement No. 614562. Rhys A. Farrer is supported by the Wellcome Trust. This project was funded in part by NIAID grant U19AI110818 to the Broad Institute. This work was also supported by independent research funded by the National Institute of Health Research (NIHR) Surgical Reconstruction and Microbiology Research Centre. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the NIH, or the Department of Health.

Figures and Tables

Figure 1. Correlation of phylogenetic and phenotypic data. Phenotypic data was superimposed onto the phylogenetic reconstruction of the nuclear genome and the data clustered into high and low value groups using a k-means clustering approach. The ability to proliferate within macrophages and to form tubular mitochondria upon engulfment are strong virulence markers. (**top left**) Mitochondrial tubularisation and yeast uptake by macrophages were correlated ($p < 0.0001$ and $p = 0.004$, respectively) with their intracellular proliferative rate (IPR). Asterisks indicate 100% bootstrap support from 1,000 replicates, and a box is used to highlight the VGIIc isolates that have shifts in phenotype.

Figure 2. Genetic changes that underlie the increased hypervirulence of *C. gattii* outbreak strains were identified by comparing EJB52 (low percent mitochondrial tubularisation and IPR) and EJB18 (high tubularisation and IPR). (**A**) Summary of all genetic differences between EJB52 and EJB18. Single base changes are shown in blue (**B**) Distance of intergenic variants between EJB52 and EJB18 to any upstream genes. (**C**) 31 genes with genetic differences were uniquely identified between EJB52 and EJB18. Genes are numerically ordered by their locus ID, and single base changes are shown in blue.

Figure 3. Phylogenetic relationships and selection of the *C. gattii* clonal subclades. (**A**) A RAxML tree with the GTRCAT model and 1,000 bootstrap replicates next to a histogram showing the number of genes with binned *dN/dS* (ω) values. (**B**) Histogram of 714 genes with $\log_{10}(\chi^2_2(2\Delta\text{Lnl}))$ values from the Branch site model (BSM) of selection in PAML. The remaining genes did not have values (e.g. due to insufficient genetic distance). The red-line is at q=0.01, which we have used as a cut-off for significance.

Table 1. *Cryptococcus gattii* strains included in this study.

Table S1. Summary of variant calling. Summary of variant calling from whole genome sequencing of 64 isolates of *Cryptococcus gattii* aligned to the nuclear genomes of VGIIa isolate R265.

Table S2. Summary of phenotypic analysis of *Cryptococcus gattii* strains. Columns show the average intracellular proliferation rates (IPR) with standard error, percent of yeast with tubular mitochondria, average percent of *C. gattii* phagocytosis by macrophages, percent of cells that were expelled without being destroyed ('vomocytosis') and the percent of macrophage depth.

Table S3. Presence/absence polymorphisms. Genes absent in VGIIc isolates EJB52 and EJB18 that are present in all of the VGIIa genomes.

Table S4. Genetic differences. All genetic differences identified between VGIIc EJB52 and VGIIc EJB18.

References

1. Fisher, M. C., Henk, D. A., Briggs, C. J., Brownstein, J. S., Madoff, L. C., McCraw, S. L. & Gurr, S. J. 2012 Emerging fungal threats to animal, plant and ecosystem health. *Nature* **484**, 186–194. (doi:10.1038/nature10947)

2. Martel, A. et al. 2014 Wildlife disease. Recent introduction of a chytrid fungus endangers Western Palearctic salamanders. *Science* **346**, 630–631. (doi:10.1126/science.1258268)

3. Singh, R. P. et al. 2011 The Emergence of Ug99 Races of the Stem Rust Fungus is a Threat to World Wheat Production. *Annu. Rev. Phytopathol.* **49**, 465–481. (doi:10.1146/annurev-phyto-072910-095423)

4. Fraser, J. A. et al. 2005 Same-sex mating and the origin of the Vancouver Island Cryptococcus gattii outbreak. *Nature* **437**, 1360–1364. (doi:10.1038/nature04220)

5. Byrnes, E. J., III et al. 2010 Emergence and Pathogenicity of Highly Virulent Cryptococcus gattii Genotypes in the Northwest United States. *PLoS Pathog* **6**, e1000850. (doi:10.1371/journal.ppat.1000850)

6. Farrer, R. A., Henk, D. A., Garner, T. W. J., Balloux, F., Woodhams, D. C. & Fisher, M. C. 2013 Chromosomal Copy Number Variation, Selection and Uneven Rates of Recombination Reveal Cryptic Genome Diversity Linked to Pathogenicity. *PLOS Genet* **9**, e1003703. (doi:10.1371/journal.pgen.1003703)

7. Raffaele, S. et al. 2010 Genome evolution following host jumps in the Irish potato famine pathogen lineage. *Science* **330**, 1540–1543. (doi:10.1126/science.1193070)

8. Datta, K. et al. 2009 Spread of Cryptococcus gattii into Pacific Northwest Region of the United States. *Emerg. Infect. Dis.* **15**, 1185–1191. (doi:10.3201/eid1508.081384)

9. Byrnes, E. J. et al. 2011 A diverse population of Cryptococcus gattii molecular type VGIII in southern Californian HIV/AIDS patients. *PLoS Pathog.* **7**, e1002205. (doi:10.1371/journal.ppat.1002205)

10. Galanis, E., MacDougall, L., Kidd, S. & Morshed, M. 2010 Epidemiology of Cryptococcus gattii, British Columbia, Canada, 1999–2007. *Emerg. Infect. Dis.* **16**, 251–257. (doi:10.3201/eid1602.090900)

11. Kidd, S. E., Guo, H., Bartlett, K. H., Xu, J. & Kronstad, J. W. 2005 Comparative gene genealogies indicate that two clonal lineages of Cryptococcus gattii in British Columbia resemble strains from other geographical areas. *Eukaryot. Cell* **4**, 1629–1638. (doi:10.1128/EC.4.10.1629-1638.2005)

12. Hagen, F. et al. 2015 Recognition of seven species in the Cryptococcus gattii/Cryptococcus neoformans species complex. *Fungal Genet. Biol. FG B* (doi:10.1016/j.fgb.2015.02.009)

13. Byrnes, E. J., Bartlett, K. H., Perfect, J. R. & Heitman, J. 2011 Cryptococcus gattii: an emerging fungal pathogen infecting humans and animals. *Microbes Infect. Inst. Pasteur* **13**, 895–907. (doi:10.1016/j.micinf.2011.05.009)

14. Byrnes, E. J., Bildfell, R. J., Frank, S. A., Mitchell, T. G., Marr, K. A. & Heitman, J. 2009 Molecular evidence that the range of the Vancouver Island outbreak of Cryptococcus gattii infection has expanded into the Pacific Northwest in the United States. *J. Infect. Dis.* **199**, 1081–1086. (doi:10.1086/597306)

15. Gillece, J. D. et al. 2011 Whole genome sequence analysis of Cryptococcus gattii from the Pacific Northwest reveals unexpected diversity. *PloS One* **6**, e28550. (doi:10.1371/journal.pone.0028550)

16. Engelthaler, D. M. et al. 2014 Cryptococcus gattii in North American Pacific Northwest: Whole-Population Genome Analysis Provides Insights into Species Evolution and Dispersal. *mBio* **5**, e01464–14. (doi:10.1128/mBio.01464-14)

17. Billmyre, R. B., Croll, D., Li, W., Mieczkowski, P., Carter, D. A., Cuomo, C. A., Kronstad, J. W. & Heitman, J. 2014 Highly Recombinant VGII Cryptococcus gattii Population Develops Clonal Outbreak Clusters through both Sexual Macroevolution and Asexual Microevolution. *mBio* **5**, e01494–14. (doi:10.1128/mBio.01494-14)

18. Feldmesser, M., Tucker, S. & Casadevall, A. 2001 Intracellular parasitism of macrophages by Cryptococcus neoformans. *Trends Microbiol.* **9**, 273–278.

19. Ma, H., Hagen, F., Stekel, D. J., Johnston, S. A., Sionov, E., Falk, R., Polacheck, I., Boekhout, T. & May, R. C. 2009 The fatal fungal outbreak on Vancouver Island is characterized by enhanced intracellular parasitism driven by mitochondrial regulation. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 12980–12985. (doi:10.1073/pnas.0902963106)

20. Bose, I., Reese, A. J., Ory, J. J., Janbon, G. & Doering, T. L. 2003 A Yeast under Cover: the Capsule of Cryptococcus neoformans. *Eukaryot. Cell* **2**, 655–663. (doi:10.1128/EC.2.4.655-663.2003)

21. Ma, H., Croudace, J. E., Lammas, D. A. & May, R. C. 2006 Expulsion of Live Pathogenic Yeast by Macrophages. *Curr. Biol.* **16**, 2156–2160. (doi:10.1016/j.cub.2006.09.032)

22. Alvarez, M. & Casadevall, A. 2006 Phagosome Extrusion and Host-Cell Survival after Cryptococcus neoformans Phagocytosis by Macrophages. *Curr. Biol.* **16**, 2161–2165. (doi:10.1016/j.cub.2006.09.061)

23. D'Souza, C. A. et al. 2011 Genome variation in Cryptococcus gattii, an emerging pathogen of immunocompetent hosts. *mBio* **2**, e00342–00310. (doi:10.1128/mBio.00342-10)

24. Farrer, R. A. et al. 2015 Genome Evolution and Innovation across the Four Major Lineages of Cryptococcus gattii. *mBio* **6**, e00868–15. (doi:10.1128/mBio.00868-15)

25. Voelz, K., Johnston, S. A., Smith, L. M., Hall, R. A., Idnurm, A. & May, R. C. 2014 'Division of labour' in response to host oxidative burst drives a fatal Cryptococcus gattii outbreak. *Nat. Commun.* **5**, 5194. (doi:10.1038/ncomms6194)

26. Staats, C. C., Kmetzsch, L., Schrank, A. & Vainstein, M. H. 2013 Fungal zinc metabolism and its connections to virulence. *Front. Cell. Infect. Microbiol.* **3**. (doi:10.3389/fcimb.2013.00065)

27. Clarke, D. L., Woodlee, G. L., McClelland, C. M., Seymour, T. S. & Wickes, B. L. 2001 The Cryptococcus neoformans STE11α gene is similar to other fungal mitogen-activated protein kinase kinase kinase (MAPKKK) genes but is mating type specific. *Mol. Microbiol.* **40**, 200–213. (doi:10.1046/j.1365-2958.2001.02375.x)

28. Urbany, C. & Neuhaus, H. E. 2008 Citrate uptake into Pectobacterium atrosepticum is critical for bacterial virulence. *Mol. Plant-Microbe Interact. MPMI* **21**, 547–554. (doi:10.1094/MPMI-21-5-0547)

29. Griffiths, E. J. et al. 2012 A defect in ATP-citrate lyase links acetyl-CoA production, virulence factor elaboration and virulence in Cryptococcus neoformans. *Mol. Microbiol.* **86**, 1404–1423. (doi:10.1111/mmi.12065)

30. Oppenheim, R. D. et al. 2014 BCKDH: the missing link in apicomplexan mitochondrial metabolism is required for full virulence of Toxoplasma gondii and Plasmodium berghei. *PLoS Pathog.* **10**, e1004263. (doi:10.1371/journal.ppat.1004263)

31. Brüggemann, H. et al. 2006 Virulence strategies for infecting phagocytes deduced from the in vivo transcriptional program of Legionella pneumophila. *Cell. Microbiol.* **8**, 1228–1240. (doi:10.1111/j.1462-5822.2006.00703.x)

32. Herrmann, V., Eidner, A., Rydzewski, K., Blädel, I., Jules, M., Buchrieser, C., Eisenreich, W. & Heuner, K. 2011 GamA is a eukaryotic-like glucoamylase responsible for glycogen- and starch-degrading activity of Legionella pneumophila. *Int. J. Med. Microbiol. IJMM* **301**, 133–139. (doi:10.1016/j.ijmm.2010.08.016)

33. Steenbergen, J. N., Shuman, H. A. & Casadevall, A. 2001 Cryptococcus neoformans interactions with amoebae suggest an explanation for its virulence and intracellular pathogenic strategy in macrophages. *Proc. Natl. Acad. Sci.* **98**, 15245–15250. (doi:10.1073/pnas.261418798)

34. Hnisz, D., Majer, O., Frohner, I. E., Komnenovic, V. & Kuchler, K. 2010 The Set3/Hos2 Histone Deacetylase Complex Attenuates cAMP/PKA Signaling to Regulate Morphogenesis and Virulence of Candida albicans. *PLOS Pathog* **6**, e1000889. (doi:10.1371/journal.ppat.1000889)

35. O'Meara, T. R. & Alspaugh, J. A. 2012 The Cryptococcus neoformans Capsule: a Sword and a Shield. *Clin. Microbiol. Rev.* **25**, 387–408. (doi:10.1128/CMR.00001-12)

36. Moyrand, F., Chang, Y. C., Himmelreich, U., Kwon-Chung, K. J. & Janbon, G. 2004 Cas3p Belongs to a Seven-Member Family of Capsule Structure

Designer Proteins. *Eukaryot. Cell* **3**, 1513–1524. (doi:10.1128/EC.3.6.1513-1524.2004)

37. Yang, Z. 2007 PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591. (doi:10.1093/molbev/msm088)

38. Chiranand, W., McLeod, I., Zhou, H., Lynn, J. J., Vega, L. A., Myers, H., Yates, J. R., Lorenz, M. C. & Gustin, M. C. 2008 CTA4 transcription factor mediates induction of nitrosative stress response in Candida albicans. *Eukaryot. Cell* **7**, 268–278. (doi:10.1128/EC.00240-07)

39. Kalayoglu, M. V., Indrawati, null, Morrison, R. P., Morrison, S. G., Yuan, Y. & Byrne, G. I. 2000 Chlamydial virulence determinants in atherogenesis: the role of chlamydial lipopolysaccharide and heat shock protein 60 in macrophage-lipoprotein interactions. *J. Infect. Dis.* **181 Suppl 3**, S483–489. (doi:10.1086/315619)

40. Dobbin, C. A., Smith, N. C. & Johnson, A. M. 2002 Heat shock protein 70 is a potential virulence factor in murine toxoplasma infection via immunomodulation of host NF-kappa B and nitric oxide. *J. Immunol. Baltim. Md 1950* **169**, 958–965.

41. Brown, J. S., Gilliland, S. M. & Holden, D. W. 2001 A Streptococcus pneumoniae pathogenicity island encoding an ABC transporter involved in iron uptake and virulence. *Mol. Microbiol.* **40**, 572–585.

42. Rodriguez, G. M. & Smith, I. 2006 Identification of an ABC transporter required for iron acquisition and virulence in Mycobacterium tuberculosis. *J. Bacteriol.* **188**, 424–430. (doi:10.1128/JB.188.2.424-430.2006)

43. Angelini, C., Bello, L., Spinazzi, M. & Ferrati, C. 2009 Mitochondrial disorders of the nuclear genome. *Acta Myol.* **28**, 16–23.

44. Farrer, R. A., Henk, D. A., MacLean, D., Studholme, D. J. & Fisher, M. C. 2013 Using false discovery rates to benchmark SNP-callers in next-generation sequencing projects. *Sci. Rep.* **3**, 1512. (doi:10.1038/srep01512)

45. Hu, G., Chen, S. H., Qiu, J., Bennett, J. E., Myers, T. G. & Williamson, P. R. 2014 Microevolution during serial mouse passage demonstrates FRE3 as a virulence adaptation gene in Cryptococcus neoformans. *mBio* **5**, e00941–00914. (doi:10.1128/mBio.00941-14)

46. Wartenberg, A. et al. 2014 Microevolution of Candida albicans in Macrophages Restores Filamentation in a Nonfilamentous Mutant. *PLOS Genet* **10**, e1004824. (doi:10.1371/journal.pgen.1004824)

47. Brunke, S. et al. 2014 One small step for a yeast--microevolution within macrophages renders Candida glabrata hypervirulent due to a single point mutation. *PLoS Pathog.* **10**, e1004478. (doi:10.1371/journal.ppat.1004478)

48. Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q. & Nordborg, M. 2012 A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* **44**, 1066–1071. (doi:10.1038/ng.2376)

49. Sironi, M., Cagliani, R., Forni, D. & Clerici, M. 2015 Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nat. Rev. Genet.* **16**, 224–236. (doi:10.1038/nrg3905)

50. Kryazhimskiy, S. & Plotkin, J. B. 2008 The Population Genetics of dN/dS. *PLoS Genet.* **4**. (doi:10.1371/journal.pgen.1000304)

51. Voelz, K. et al. 2013 Transmission of Hypervirulence Traits via Sexual Reproduction within and between Lineages of the Human Fungal Pathogen Cryptococcus gattii. *PLoS Genet.* **9**. (doi:10.1371/journal.pgen.1003771)

52. Springer, D. J. et al. 2014 Cryptococcus gattii VGIII Isolates Causing Infections in HIV/AIDS Patients in Southern California: Identification of the Local Environmental Source as Arboreal. *PLoS Pathog.* **10**. (doi:10.1371/journal.ppat.1004285)

53. Li, H. 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio*

54. Li, H. et al. 2009 The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079. (doi:10.1093/bioinformatics/btp352)

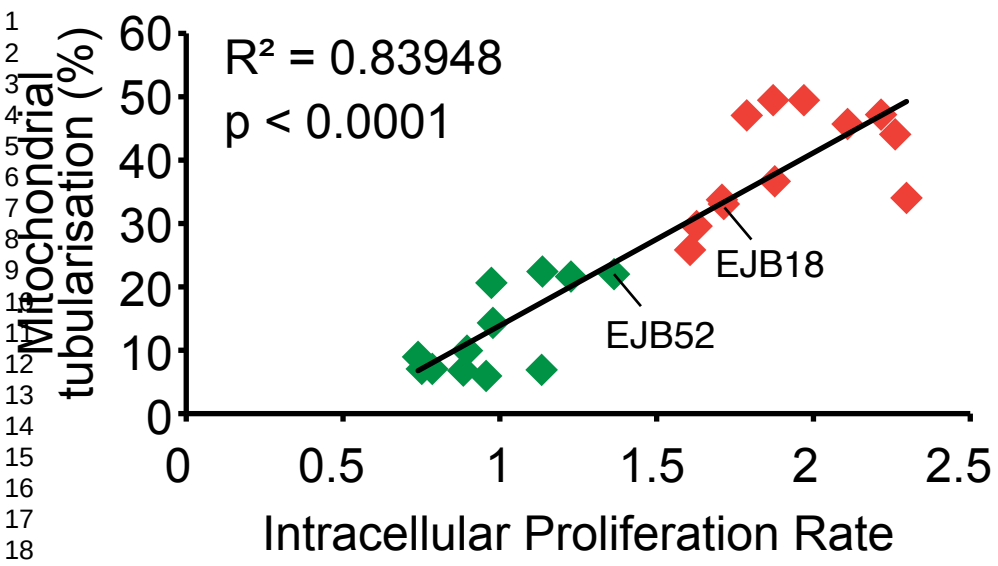
55. McKenna, A. et al. 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303. (doi:10.1101/gr.107524.110)

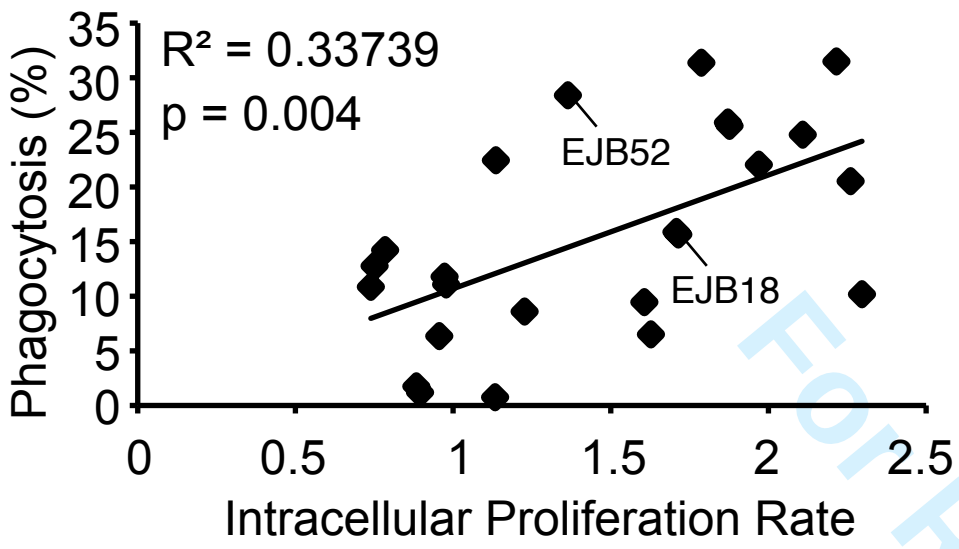
56. In press. Picard Tools - By Broad Institute.

57. Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. 2007 Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* **2**, 953–971. (doi:10.1038/nprot.2007.131)

58. Yang, Z. & Nielsen, R. 2000 Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43.

Strain Type	Strain	Origin	est. Lat.	est. Long.	Source	Mating Type
VGI	B7488	USA, Oregon	43.8	-120.5	Clinical	alpha
VGI	E566	South Australia	-25.5	134.0	eucalyptus tree	a
VGI	EJB2	USA North Carolina, with a history throughout the San Francisco	46.0	-121.0	Clinical	alpha
VGI	NT-10	Australia	-25.5	134.0	Clinical	alpha
VGI	Ru294	South Africa	-31.0	23.7	unknown tree	alpha
VGI	WM179	Sydney, Australia, 1993	-33.9	151.1	Clinical	alpha
VGI	WM276	Australia	-25.5	134.0	Environmental	alpha
VGII	2001/935	Senegal	14.2	-14.4	Clinical	alpha
VGII	99/473	Caribbean Islands	20.8	-77.6	Clinical	alpha
VGII	CBS10089	Greece	39.5	21.8	Clinical	alpha
VGII	CBS1930	Aruba, Caribbean Sea	12.5	-70.0	Goat	a
VGII	CBS8684	Uruguay	-32.9	-56.0	Environmental (Wasp nest)	alpha
VGII	ICB180	Sao Paulo, Brazil	-9.5	-55.8	Environmental (Eucalyptus tree)	alpha
VGII	ICB184	Piaui, Brazil	-9.5	-55.8	Environmental (Tree)	alpha
VGII	LA362	Brazil,Jaboticabal	-21.3	-48.3	Animal (Parrot lier?)	alpha
VGII	LMM265	Brazil	-9.5	-55.8	Clinical	alpha
VGII	MMRL2647	Caribbean Islands	-25.5	134.0	Clinical	alpha
VGIIa	B7395	USA,Washington	38.9	-77.0	Dog	alpha
VGIIa	B7422	USA, Oregon	43.8	-120.5	Cat	alpha
VGIIa	B7467	USA, Washington	38.9	-77.0	Porpoise	alpha
VGIIa	B8577	Canada, British Columbia	53.9	-127.6	Environmental	alpha
VGIIa	B8849	USA,Oregon	43.8	-120.5	Environmental	alpha
VGIIa	CA1014	USA	46.0	-121.0	Clinical	alpha
VGIIa	CBS10485	Canada, Vancouver Island	49.7	-125.2	Clinical (Danish tourist)	alpha
VGIIa	CDCF2932	Canada, British Columbia, Kelowna	49.9	-119.5	Clinical (Immunocompetent patie	alpha
VGIIa	CDCF3016	Canada, shores island close to Vancouver Island	49.7	-125.2	Animal (Dead wild Dall's	alpha
VGIIa	CDCR271	Canada, British Columbia, Nanoose Bay	49.3	-124.2	Clinical (Immunocompetent male)	alpha
VGIIa	ENV152	Canada, Vancouver Island, Provincial Park, Rathrevor Beach	49.3	-124.3	Environmental (Alder tree)	alpha
VGIIa	R265	Canada, British Columbia, Duncan	48.8	-123.7	Clinical	alpha
VGIIb	B7394	USA, Washington	38.9	-77.0	Cat	alpha
VGIIb	B7735	USA, Oregon	43.8	-120.5	Clinical	alpha
VGIIb	B8554	USA, Oregon	43.8	-120.5	Dog	alpha
VGIIb	B8828	USA, Washington	38.9	-77.0	Porpoise	alpha
VGIIb	Ram5	Australia	-25.5	134.0	Clinical	alpha
VGIIc	B6863	USA, Oregon	43.8	-120.5	Clinical	alpha
VGIIc	B7390	USA, Idaho	44.2	-114.8	Clinical	alpha
VGIIc	B7432	USA, Oregon	43.8	-120.5	Clinical	alpha
VGIIc	B7466	USA, Oregon	43.8	-120.5	Cat	alpha
VGIIc	B7737	USA, Oregon	43.8	-120.5	Clinical	alpha
VGIIc	B8571	USA, Washington	38.9	-77.0	Clinical	alpha
VGIIc	B8838	USA,Washington	38.9	-77.0	Clinical	alpha
VGIIc	B8843	USA,Oregon	43.8	-120.5	Clinical	alpha
VGIIc	EJB18	USA,Oregon	43.8	-120.5	Clinical	alpha
VGIIc	EJB52	USA,Oregon	43.8	-120.5	Clinical	alpha
VGIIIa	78-1-S3A	Los Angeles, California, USA, 2011	34.0	-118.3	Environmental	alpha
VGIIIa	97/433	Mexico	23.4	-101.7	Clinical	alpha
VGIIIa	B8212	USA, Oregon	43.8	-120.5	unknown	alpha
VGIIIa	BHPP3-S1A	Los Angeles, California, USA, 2012	34.0	-118.3	Environmental, Soil	alpha
VGIIIa	CA1053	California, USA	36.5	-119.7	Clinical	alpha
VGIIIb	CA1280	USA	46.0	-121.0	Clinical	alpha
VGIIIa	CA1308	California, USA	36.5	-119.7	Clinical	alpha
VGIIIb	CA1508	California, USA	36.5	-119.7	Clinical	a
VGIIIa	CA1873	USA	46.0	-121.0	Clinical	a
VGIIIa	MCP-1A	Los Angeles, California, USA, 2011	34.0	-118.3	Environmental	alpha
VGIIIb	MCPR1-S1B	Los Angeles, California, USA, 2012	34.0	-118.3	Environmental, Soil	a
VGIIIb	NIH312	California, USA	36.5	-119.7	Clinical	alpha
VGIIIb x VGIIx	Progeny5	NA	NA	NA	NA	alpha
VGIIIa	WM11_139	USA, 2011	43.8	-120.5	Veterinary	a
VGIIx	CBS10090_Bir	Greece	39.5	21.8	Clinical	a
VGIIx	CBS10090_Bro	Greece	39.5	21.8	Clinical	a
VGIIx	LA55	NE region of Piaui, Brazil	-9.5	-55.8	Clinical	a
VGIV	CBS10101	South Africa	-31.0	23.7	King Cheetah	alpha
VGIV	IND107	India	22.4	78.9	Clinical	alpha
VGIV	WM779	South Africa, 1994	-31.0	23.7	Veterinary	alpha

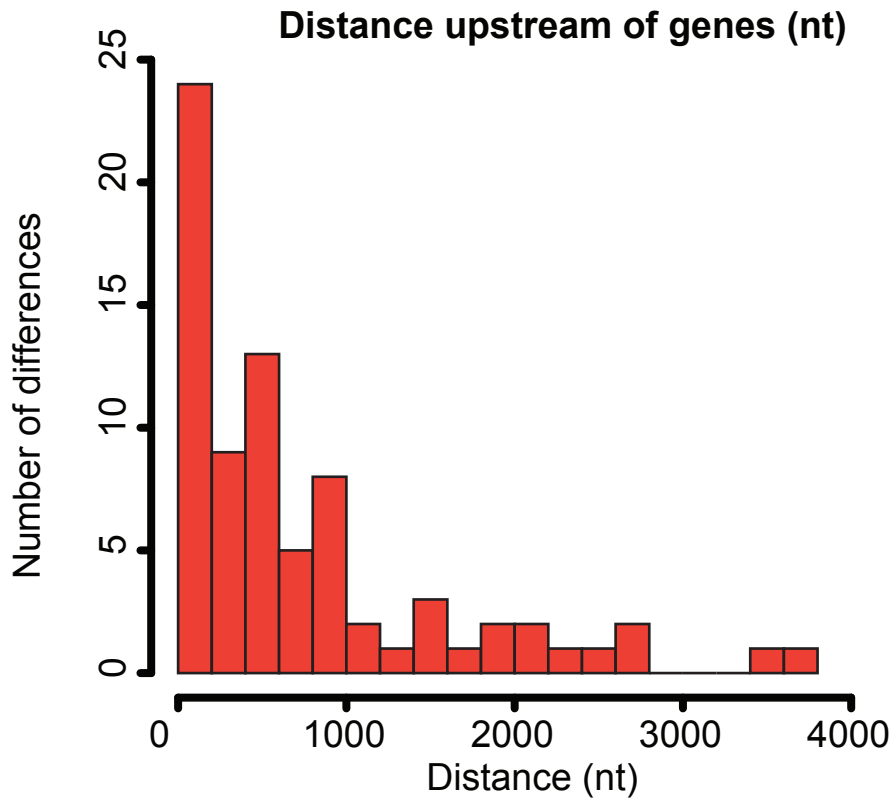




A)

Category	EJB52	EJB18	Total differences (nt)	Percent of Genome
Change	Reference	SNP	12	6.95E-05
	SNP	Reference	20	1.16E-04
	SNP	Different SNP	2	1.16E-05
	Insertion	Dif. Insertion	7	4.06E-05
	Insertion	Reference	35	2.03E-04
	Reference	Insertion	25	1.45E-04
	Deletion	Dif. Deletion	3	1.74E-05
	Deletion	Insertion	1	5.79E-06
	Deletion	Reference	28	1.62E-04
	Reference	Deletion	20	1.16E-04
Fixed	Reference	Reference	16,967,822	98.30
	SNP	SNP	54,250	0.31
	Insertion	Insertion	2,568	0.01
	Deletion	Deletion	3,096	0.02
Ambiguous	Reference	Ambiguous	19,660	0.11
	SNP	Ambiguous	319	1.85E-03
	Ambiguous	SNP	167	9.68E-04
	Insertion	Ambiguous	18	1.04E-04
	Deletion	Ambiguous	29	1.68E-04
	Ambigious	Reference	6,615	0.04
	Ambigious	Ambigious	206,065	1.19

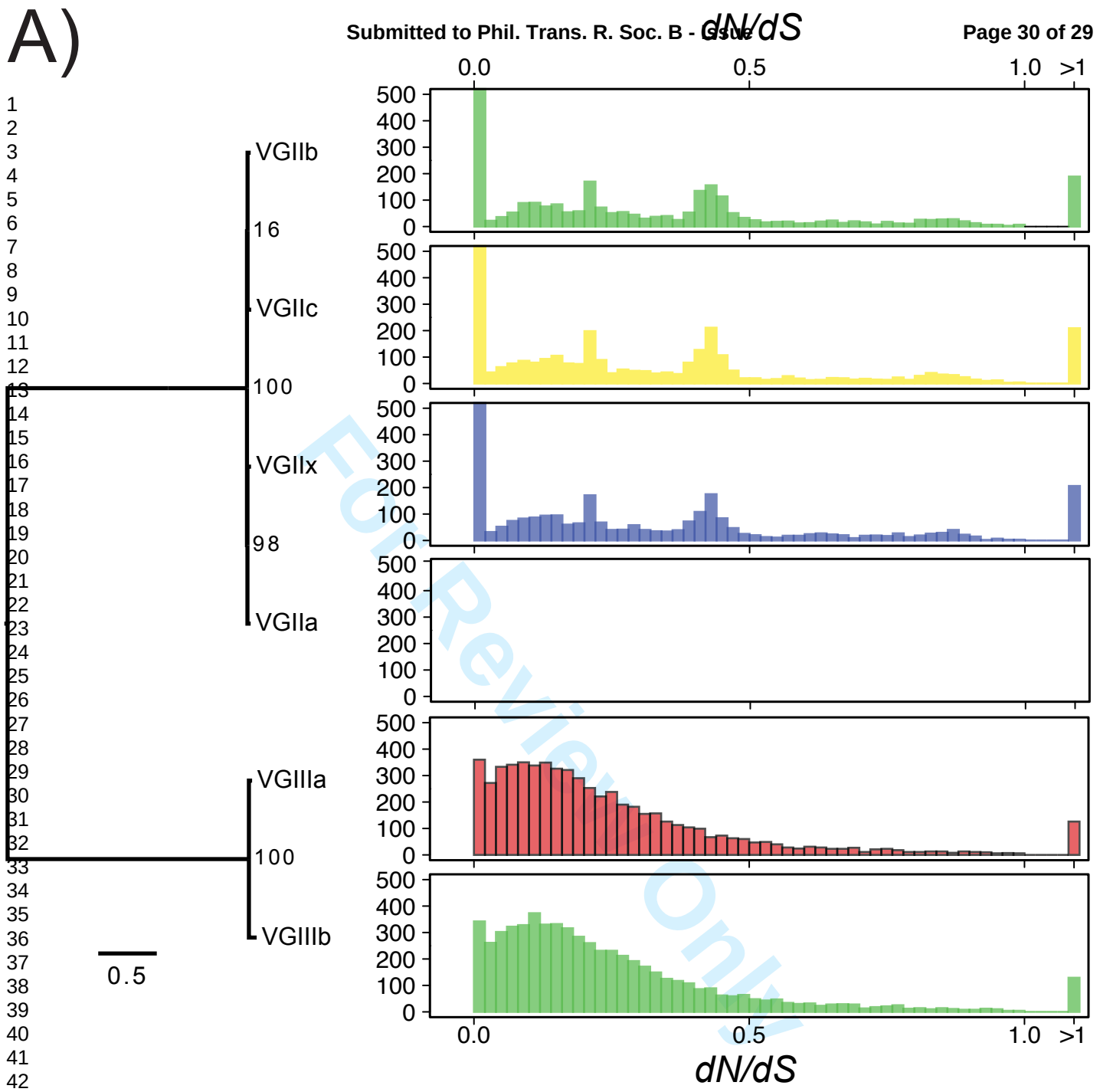
B)



C)

Gene ID	EJB52	EJB18	Annotation	PFAM domains
CNBG_0001	SNP (Non-synonymous)	Reference	Hypothetical protein	Late exocytosis, associated with Golgi transport; Phosphate metabolism
CNBG_0232	Insertion (Non-frameshift)	Reference	Hypothetical protein	SUZ domain
CNBG_0580	SNP (Non-synonymous)	Reference	Nup155	Nup133; Non-repetitive/WGA-negative nucleoporin
CNBG_1027	Reference	SNP (Non-synonymous)	Cerevisin	Subtilase family; Peptidase inhibitor I9
CNBG_1088	Insertion (Frameshift)	Reference	Ligase	N/A
CNBG_1192	Deletion (Non-frameshift)	Reference	Hypothetical protein	SNF5 / SMARCB1 / INI1
CNBG_1482	Insertion (Non-frameshift)	Reference	Hypothetical protein	N/A
CNBG_1665	Deletion (Non-frameshift)	Reference	HSE1	N/A
CNBG_1847	Deletion (Frameshift)	Insertion (Frameshift)	Histone deacetylase RPD3	Histone deacetylase domain
CNBG_2344	Reference	SNP (Non-synonymous)	Sterol 3β-glucosyltransferase	UDP-glucoronosyl and UDP-glucosyl transferase (GTF); GTF 28
CNBG_2480	Reference	SNP (Non-synonymous)	Efflux protein EncT	Major Facilitator Superfamily
CNBG_2491	Deletion (Non-frameshift)	Reference	Hypothetical protein	Arrestin (or S-antigen), C-terminal domain
CNBG_2491	Reference	Insertion (Non-frameshift)	Hypothetical protein	Arrestin (or S-antigen), C-terminal domain
CNBG_2628	Reference	Deletion (Non-frameshift)	Hypothetical protein	N/A
CNBG_2696	Insertion (Non-frameshift)	Reference	Chaperone	SRP40, C-terminal domain; LisH
CNBG_2828	Reference	SNP (Non-synonymous)	PEPCK	Phosphoenolpyruvate carboxykinase
CNBG_3379	Reference	SNP (Synonymous)	α-1,6-mannosyltransferase	Alg9-like mannosyltransferase family
CNBG_3803	SNP (Non-synonymous)	Reference	Hypothetical protein	N/A
CNBG_3873	SNP (Synonymous)	Reference	Histone deacetylase 6/10	Histone deacetylase domain; Arb2 domain
CNBG_3995	SNP (Non-synonymous)	Reference	Carboxypeptidase A4	Zinc carboxypeptidase
CNBG_4063	Insertion (Non-frameshift)	Reference	Hypothetical protein	N/A
CNBG_4812	SNP (Synonymous)	Reference	ADP,ATP carrier protein	Mitochondrial carrier protein
CNBG_5309	Deletion (Non-frameshift)	Dif. Del. (Non-frameshift)	Cytoplasmic protein	PSP1 C-terminal conserved region
CNBG_5312	Reference	SNP (Non-synonymous)	Glucoamylase	N/A
CNBG_5651	Reference	Insertion (Non-frameshift)	Hypothetical protein	N/A
CNBG_5749	SNP (Non-synonymous)	Reference	Hypothetical protein	WD domain, G-beta repeat domain
CNBG_5891	Insertion (Non-frameshift)	Reference	Hypothetical protein	WD domain, G-beta repeat domain
CNBG_5891	Reference	Deletion (Non-frameshift)	Hypothetical protein	WD domain, G-beta repeat domain
CNBG_6031	SNP (Non-synonymous)	Reference	Hypothetical protein	2'-5' RNA ligase superfamily; Cyclic phosphodiesterase-like
CNBG_6184	Insertion (Non-frameshift)	Reference	Hypothetical protein	N/A
CNBG_6204	Insertion (Non-frameshift)	Reference	Hypothetical protein	N/A
CNBG_9229	SNP (Non-synonymous)	Reference	Hypothetical protein	RhoGEF domain; Variant SH3 domain x2; SH3 domain; WH2 motif
CNBG_9578	Reference	Deletion (Non-frameshift)	Hypothetical protein	N/A

A)



B)